Systematics and symmetry in molecular phylogenetic modelling

IOP Conf. Series: Journal of Physics: Conf. Series **1194** (2019) 012056 doi:10.1088/1742-6596/1194/1/012056

Systematics and symmetry in molecular phylogenetic modelling.

Peter D Jarvis

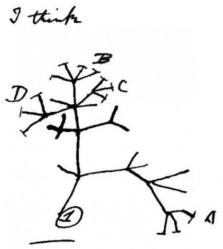
School of Natural Sciences (Mathematics and Physics), University of Tasmania, Private Bag 37 GPO, Hobart Tas 7001 Australia

E-mail: Peter.Jarvis@utas.edu.au

Abstract. Probabilistic models of mathematical phylogenetics have been intensively used in recent years in biology as the cornerstone of methods to infer and reconstruct the ancestral relationships between species. We bring a lens of mathematical physics to bear on the formulation of the theoretical models, focussing on the applicability of group and representation theory to guide model specification and to exploit the multilinear setting of the models in the presence of underlying symmetries. We focus on aspects of multipartite entanglement which are shared between descriptions of quantum states on the physics side, and the multi-way tensor probability arrays arising in phylogenetics. We give examples of entanglement invariants (Markov invariants) for the case of quartets, and DNA data; and compare and contrast the rings of quantum/Markov invariants for the three and four qubit case/binary triplet and quartet cases, respectively.

1. Inferring phylogenies: the challenge of ancestral reconstruction

The task of phylogenetics is to construct ancestral relationships (phylogenies), inferred by analyzing statistical data, collected for various (morphological or genotypic) kinds of characters or traits, possessed by selected groups of biological organisms (taxa). The first modern phylogenetic tree ever drawn, from Charles Darwin's 1837 notebook [1], considerably earlier than the eventual publication of *The Origin of Species*, confronts head on the question of human evolutionary origins:



Darwin, Wallace and the early proponents of evolution had no access to genetic information (Mendel's work was contemporaneous but only rediscovered after another forty years), let alone molecular data. Based on comparison of homologous morphological features, the natural principle for ancestral reconstruction was then, and still is in many studies, that of *parsimony* – the path from a common ancestor to an extant organism should be the *most conservative*, exhibiting fewest changes.

	P	G	H	C	B
P	□	12	15	15	10
G		□	9	8	15
H			□	3	10
C				□	17
B					□

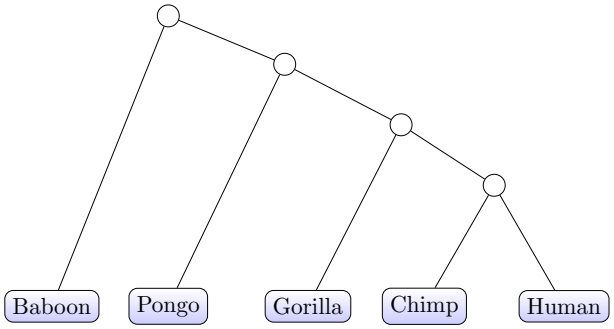


Figure 1. A mock-up of Darwin's problem of great ape and human ancestry in terms of parsimony: pairwise difference scores taken as distance measures, are used to create a tree .

As a simple illustration, suppose we have pairwise comparisons of 'degree of similarity' between 5 species ('taxonomic units') $\{P, G, H, C, B\}$ as in figure 1: Clearly, the pairs with the smallest scores should be assigned to an ancestral tree in close proximity, in this case $P-G$, $G-H$ with $G-C$, and notably $H-C$. If G is identified as an outgroup of the sibling pair ('cherry') $H-C$, the placement of P and B as further outgroups is consistent . This (artificial) example, resulting in figure 1, is of course a mock-up[1] of the evolutionary situation of the great apes – {Pongo, Gorilla, Human, Chimpanzee, Baboon} – Darwin's original problem. In complicated cases where it is no longer apparent how to proceed by hand, it turns out that there is an efficient greedy algorithm which is able to resolve ties, and produce an optimal tree under *maximum parsimony* [2].

2. Molecular phylogenetics model building

The era of molecular sequence-based phylogenetic modelling was ushered in as a result of dramatic advances from the 'fifties and 'sixties with the unravelling of the genetic code. Work of Pauling, Zuckerkandl and others on the sequence and structure of families of relatively small proteins such as haemoglobin and myoglobin had demonstrated the fact of molecular evolution; subsequent analysis, in the nucleic acid case, of the ubiquitous ribosomal structures of the cell and their attendant transfer molecules, laid bare the evidence that these systems were at a fundamental level subject to random substitutions, in the vast majority of cases benign, but whose presence in sequence data could hold powerful information about ancestry. Indeed, the *neutral theory of evolution* [3] is exploited to construct probabilistic, parametrised models of ancestral relationships, which can be fitted to the molecular sequence data, whose ready availability since the 80's has brought about a revolution in the potential for ancestral reconstruction, whether the data come from sequences of proteins (20 amino acids), or nucleic acids (4 DNA/RNA base letters). In this section we elaborate the construction of such models via a tensorial, multilinear algebraic perspective. This leads naturally to considerations of group actions and symmetries, and (in the next section) aspects of invariant theory and entanglement in this setting.

The reality of modern phylogenetic analysis is shown in following figure 2. Beginning with a *sequence alignment*, that is, a string of N letters (sequence length), from an alphabet of size K (number of characters), for each of L rows (number of taxa), the aim is to produce a corresponding *phylogenetic tree*:

[1] Adapted from `http://www.calvin.edu/~rpruim/talks/SC11/2011-06/SC11-Calvin-Rendon/`

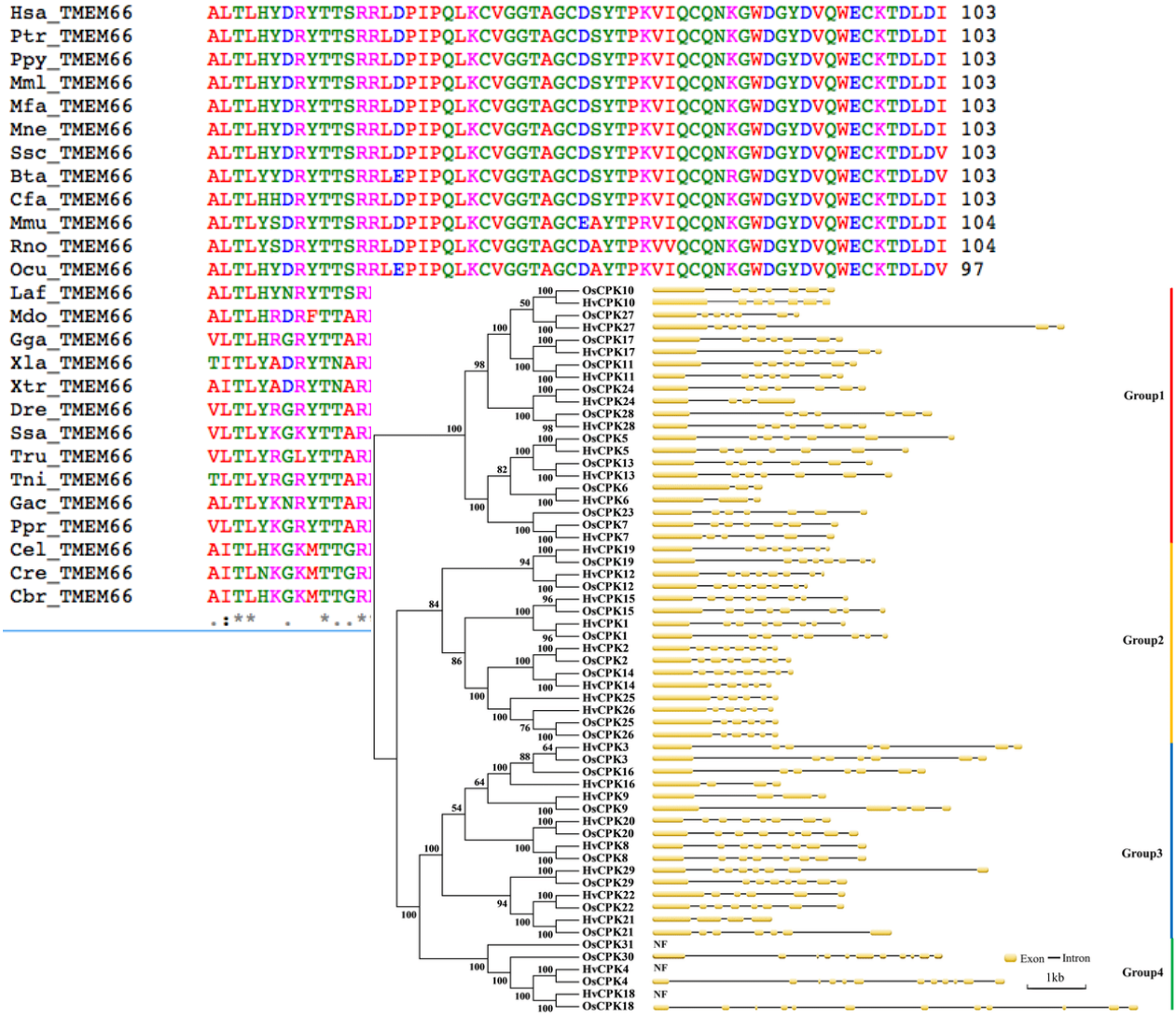


Figure 2. The protein sequence alignment [4] refers to a family of membrane proteins, while the tree [5] (unrelated data) assigns kinase proteins of rice and barley to four main clades (subtrees), also annotated for each gene with the pattern of introns and exons involved in post-transcriptional splicing. These samples exemplify the richness of phylogenetics: the membrane protein alignment indicates molecular evolution across species, while the rice-barley tree manifests sub-specific sructure.

The task of phylogenetics in reducing sequence data to ancestral relationships is the following [6, 7]:

Phylogenetic data and inference:

(a) Convert the data into an array of K^L *pattern frequencies* (the relative frequency of occurrence of each pattern of characters at each of the N sites), collectively representing an L-way contingency table T_{align} ;
(b) Choose a tree $\mathcal{T}$ with L leaves;
(c) Given some set of parameters $\underline{\alpha}$ representing substitution probabilities between character states, use these and $\mathcal{T}$ to build a *probability tensor* under a standard algorithm;

$$P_{\mathcal{T}}(\underline{\alpha})^{i_1 i_2 \cdots i_L} , \quad i_1, i_2, \cdots i_L = 1, 2, \cdots, K ;$$

(d) Assuming the data is an i.i.d. *multinomial distribution* sample

$$T_{\text{align}} \cong \mathsf{Mult}(K^L, P_{\mathcal{T}}(\underline{\alpha})) ,$$

use statistical inference to produce the optimal tree $\mathcal{T}^*$ *and model parameters* $\underline{\alpha}^*$.

□

Given this, the multilinear setting is now clear. The required phylogenetic tensor is an object in the (probability simplex in) the space $\otimes^L(\mathbb{C}^K)$. There is also a natural correspondence with 'second quantization' – machinery is needed to go from a vector (the root distribution, say π) to an element of the L-fold tensor product. A formal description uses the following ingredients:

Branching (phylogenesis)
The splitting (branching) operator δ (a comultiplication) defined in the natural basis[2], by $\delta(e_i) := e_i \otimes e_i$, and extended linearly:

$$\delta(\sum\nolimits_i \pi^i e_i) = \sum\nolimits_i \pi^i e_i \otimes e_i .$$

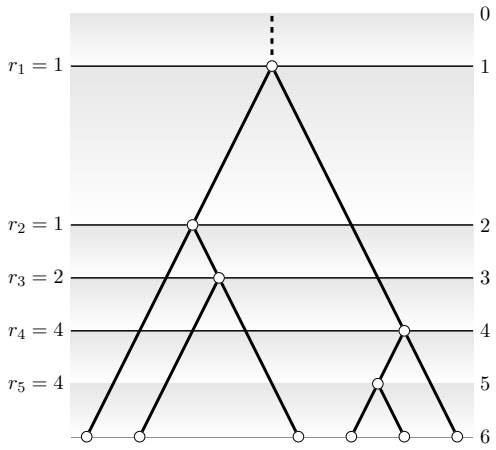
Phyletic evolution (anagenesis)
Stochastic evolution on each edge is governed by independent Markov processes which generate an appropriate product action on tensors via $M \otimes M' \otimes M'' \cdots$.

□

At the level of underlying stochastic processes, the role of δ is to generalize from a Markov chain to a tree [7], in that states on edges now have joint dependence due to a common source node ('conditional independence'). Note that the M-action preserves a *linear* form: $\eta(M \cdot \pi) = \eta(\pi) = \sum_i \eta_i \pi^i := \sum_i \pi^i$ – so the M are *invertible unit column-sum* (entries positive) matrices. We call the corresponding complex group $\cong GL_1(K, \mathbb{C})$ the *Markov group* (isomorphic to the complex affine group in dimension $K-1$). Concretely, a candidate phylogenetic tensor is constructed as follows:

[2] 'Copying', 'entangling', 'cloning', $|i\rangle \mapsto |i, i\rangle$ if implemented directly.

- Draw the tree with numerical marks $r_1 = 1$, $r_2 \in \{1,2\}, \cdots$, $r_{L-1} \in \{1,2,\cdots,L-1\}$ for $(L-1)$ bifurcations, and choose $(L-1)$ pairs of transition matrices for the sibling edges ($2L-2$ edges including pendant leaves);
- Grow the tensor starting from the root π by iterating the insertion of $M_r \otimes M'_r \circ \delta^{(r)}$, $r = 1, 2, \cdots, L-1$.



Proceeding algebraically, the phylogenetic tensor is constructed incrementally, viz.

$$P^{(1)} = \pi\,, \quad P^{(2)} = M_1 \otimes M'_1 \circ \delta \circ P^{(1)}\,, \quad P^{(3)} = M_2 \otimes M'_2 \otimes \mathbb{I} \circ \delta \otimes \mathsf{Id} \circ P^{(2)}\,, \cdots\,,$$

until, at the final step, the fully-formed tensor is $P_{\mathcal{T}} \equiv P^{(6)}$, namely

$$P^{(6)} = \mathbb{I}^3 \otimes M_5 \otimes M'_5 \otimes \mathbb{I} \circ \mathsf{Id}^3 \otimes \delta \otimes \mathsf{Id} \circ P^{(5)}\,.$$

The bias-variance trade-off in statistical inference means that it is often inappropriate to use the most general transition matrix with arbitrary parameters. For DNA models specific, restricted choices are often used based on suppositions about substitution rates, to avoid over-parametrization and reflect biological realities. A natural criterion for heterogeneous models is that of *multiplicative closure*: if a node between edges in a tree is omitted, then there should be an interpolating transition matrix belonging to the same class, that is (introducing rate matrices Q where $M = e^Q$), $e^{\overline{Q}} = e^{Q_1} e^{Q_2}$. This is a surprisingly strong condition. With the use of the Baker-Campbell Hausdorff formula, a sufficient condition for the existence of a matrix $\overline{Q}$ in the same class (but not necessarily for it to be a valid rate matrix) is that the model rate matrices form a Lie algebra. Further examination of the behaviour of the rate parameters under label permutations leads to the *Lie-Markov* hierarchy [8, 9, 10] which includes (with some notable exceptions) most of the phylogenetic models in common use.

Example: In the *strand symmetric model*, the transition matrix is

$$M^{SSM} = \begin{pmatrix} M_{\mathtt{AA}} & M_{\mathtt{AC}} & M_{\mathtt{AG}} & M_{\mathtt{AT}} \\ M_{\mathtt{CA}} & M_{\mathtt{CC}} & M_{\mathtt{CG}} & M_{\mathtt{CT}} \\ M_{\mathtt{GA}} & M_{\mathtt{GC}} & M_{\mathtt{GG}} & M_{\mathtt{GT}} \\ M_{\mathtt{TA}} & M_{\mathtt{TC}} & M_{\mathtt{TG}} & M_{\mathtt{TT}} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{a} & b & c & d \\ e & \mathbf{f} & g & h \\ h & g & \mathbf{f} & e \\ d & c & b & \mathbf{a} \end{pmatrix}$$

with $\mathbf{a} = 1-d-e-h$ $\mathbf{f} = 1-b-c-g$, consistent with Crick-Watson pairing $\mathtt{G} \leftrightarrow \mathtt{C}$ and $\mathtt{A} \leftrightarrow \mathtt{T}$, and stationary frequencies $\pi_{\mathtt{G}} = \pi_{\mathtt{C}}$, $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ (c.f. Chargaff's rule). The SSM has $\mathbb{Z}_2 \wr \mathbb{Z}_2$ permutation symmetry with respect to rearrangement of its nucleotides [11]. □

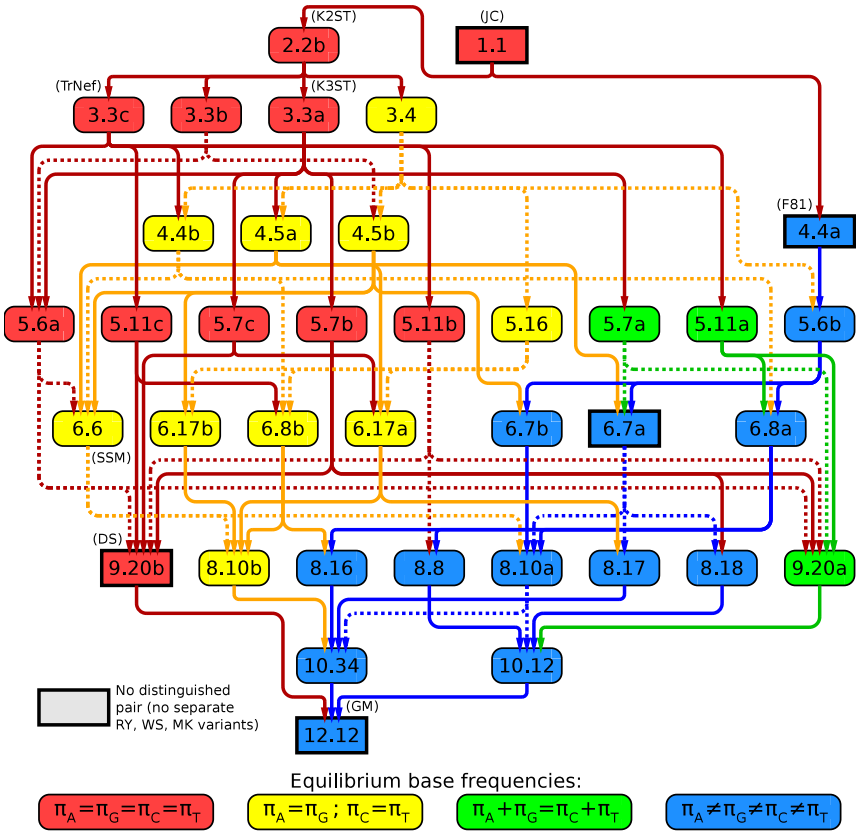


Figure 3. Diagram depicting the hierarchy of Lie-Markov models with nucleotide permutation symmetry $\mathbb{Z}_2 \wr \mathbb{Z}_2$ and their interrelationships. Common models are identified by acronyms (for citations see [12]), with the strand symmetric model being 6.6. The labelling $d.g$ gives the dimension d of the Lie algebra, and the number g of extremal rays in the stochastic cone spanning the rate matrices of the model. Figure courtesy of Michael Woodhams [10].

3. Entanglement

The notion of independence of random events is ubiquitous in statistics and stochastic modelling. Familiar instances are Hardy-Weinberg equilibrium, and linkage (dis)equilibrium, in population genetics. The multilinear, tensor framework for phylogenetics modelling is a natural setting for incorporating statistical independence; indeed, we have already seen above that the underlying framework for constructing phylogenetic trees is that of *conditional* independence of (random variables describing sibling states) on a common ancestor. By the same token, given a phylogenetic tree and phylogenetic tensor, additional edge evolution (anagenetic, without further speciation), engenders the Markov action

$$P \to P' := M_1 \otimes M_2 \otimes \cdots \otimes M_L \cdot P$$

– while the speciation or splitting (comultiplication) operator δ generates correlations, the joint stationary distribution (after sufficient time) will correspond to *rank 1* tensors $P = \pi_1 \otimes \pi_2 \otimes \cdots \otimes \pi_L$, the tensor product of L individual stationary vectors. The correspondence at this level between the theoretical descriptions of multipartite quantum systems, on the one hand, and multi-taxon phylogenetic distributions, is striking: in the former, we have complex

IOP Conf. Series: Journal of Physics: Conf. Series **1194** (2019) 012056 doi:10.1088/1742-6596/1194/1/012056

Table 1. Table comparing and contrasting multipartite quantum systems, and multitaxon phylogenetic systems, as to state space, dimension, and symmetry group acting locally (see text and [13]).

	1 particle	L particle	local symmetry
quantum (pure states)	$\mathbb{C}^K$	$\otimes^L\mathbb{C}^K$	$\times^L U(K)$
quantum (mixed states)	$\mathbb{C}^K \otimes \mathbb{C}^{K*}$	$\otimes^L\mathbb{C}^{K^2}$	$\times^L U(K)$
stochastic	$\mathbb{C}^K$	$\otimes^L\mathbb{C}^K$	$\times^L GL_1(K)$
stochastic Lie-Markov (symmetry G)	$\mathbb{C}^K$	$\otimes^L\mathbb{C}^K$	$\times^L G$, $G < GL_1(K)$

Hilbert spaces and a local unitary group action (or adjoint action for mixed states)[3], and in the latter, we have the action of the complex local Markov group as local symmetry.

This is precisely the mathematical setting of classical invariant theory [13]. Just as, in the quantum case, the local unitary invariants are the ingredients for measures of entanglement, a natural question for the phylogenetics case is to identify *Markov invariants* – quantities invariant under the local group action up to scaling, that may be able to provide model-independent information about the underlying tree. Such invariants include the following [13, 14]:

Det ($L = 2$, degree K)

$$\mathrm{Det}(P) = \sum P^{i_1 j_1} P^{i_2 j_2} \cdots P^{i_K j_K} \varepsilon_{i_1 i_2 \cdots i_K} \varepsilon_{j_1 j_2 \cdots j_K}$$

and transforms as $\mathrm{Det}(P) \to \det(M_1)\det(M_2)\mathrm{Det}(P)$. Using $\log\det = \mathrm{Tr}\log$, LogDet gives a measure of total evolutionary change (time × rates) on the path between 1 & 2 on the tree. □

HyperDet ($L = 3$, binary $K = 2$, degree 4)

$$\tau(P) = (P^{111})^2(P^{222})^2 + (P^{112})^2(P^{221})^2 + (P^{121})^2(P^{212})^2 + (P^{211})^2(P^{122})^2$$
$$\cdots - 2P^{112}P^{121}P^{212}P^{221} - 2P^{112}P^{122}P^{221}P^{211} - 2P^{121}P^{122}P^{212}P^{211}$$

and transforms as $\tau(P) \to \det(M_1)^2 \det(M_2)^2 \det(M_3)^2 \tau(P)$. LogHyperDet gives a measure of total evolutionary change on the star formed by 1 & 2 & 3. □

[3] In the case of stochastic local operations, the appropriate symmetry group is in fact the complex general linear group.

Stangle ($L = 3$ $K = 2$, degree 3)

$$\begin{aligned} ST(P) &= \sum P^{i_1 i_2 i_3} P^{j_1 j_2 j_3} P^{k_1 k_2 k_3} \varepsilon_{i_1 j_1} \varepsilon_{i_2 j_2} \varepsilon_{i_3 k_3} \eta_{j_3} \eta_{k_1} \eta_{k_2} \\ &\to \det(M_1)^1 \det(M_2)^1 \det(M_3)^1 ST(P) \end{aligned}$$

□

Squangles ($L = 4$ $K = 4$, degree 5)

For four leaves, $L = 4$, called 1,2,3 & 4, and DNA sequences ($K = 4$), we find a set of 3 quintic Markov invariants $\mathbb{Q}_1, \mathbb{Q}_2, \mathbb{Q}_3$ such that $\mathbb{Q}_1 + \mathbb{Q}_2 + \mathbb{Q}_3 = 0$, and signs

	12|34	13|24	14|23
$\mathbb{Q}_1$	0	< 0	> 0
$\mathbb{Q}_2$	> 0	0	< 0
$\mathbb{Q}_3$	< 0	> 0	0

Given real (noisy) data there is a simple (signed) least squares protocol that allows a weighted optimum choice to be made [15]. There is a graphical theorem that says that the tree can be reconstructed uniquely from its quartets, and so the squangles provide a model-parameter independent means of quartet identification, and hence inference. Because of the combinatorial definition (they are sparse polynomials of degree 5 in $4^4 = 256$ variables) they each have about 50,000 terms, but they can be evaluated efficiently. □

Note that Det and HyperDet are full $GL(K)$ invariants; LogDet has long been recognised as a useful pairwise 'distance' measure between leaves, which ideally allows tree reconstruction [16]. The HyperDet (the Cayley hyperdeterminant) is known as the tangle in multipartite quantum systems; and LogHyperDet similarly provides a joint total distance for triplets [17]. The stangle and squangles[4] however are bona fide stochastic invariants, and not GL invariant: $GL_1(K) < GL(K)$.

In the quantum case, a prerequisite for the construction of measures of entanglement is the classification of all local unitary invariants, and hence the characterization of generators for the invariant ring. The stochastic transformation group, the Markov group $GL_1(K)$, is not semi-simple, but the algebraic structure can be determined enumeratively. Table 2 gives the 'empirical' situation for the case of multi-qubit systems and the analogous multi-taxon binary character systems in phylogenetics.

In particular, we have the following results for binary triplets [21]:

Fundamental invariants

$\Phi(P)$: the probability mass, $\Phi(P) := \eta_i \eta_j \eta_k P^{ijk}$.
$\tau(P)$: the tangle, as above.
$\mathrm{Det}_{ab}(P)$: $\mathrm{Det}_{12}(P) := \det(\eta_k P^{ijk})$; similarly $\mathrm{Det}_{13}(P)$, $\mathrm{Det}_{23}(P)$.
$ST(P)$: the stangle, as above.

Syzygies There is a degree 6 identity,

$$\Phi^3 ST - \Phi^2 \tau + 4 D_{12} D_{23} D_{31} = 0$$

– so the Molien series for the free ring composed from fundamental invariants, gets modified:

$$\begin{aligned} h_{free} &= \frac{1}{(1-q)(1-q^2)^3(1-q^3)(1-q^4)} \\ \Rightarrow \quad h(q) &= \frac{(1-q^6)}{(1-q)(1-q^2)^3(1-q^3)(1-q^4)} \equiv \frac{(1+q^3)}{(1-q)(1-q^2)^3(1-q^4)} \end{aligned}$$

[4] The nomenclature '$s\cdots$x$\cdots$angle' incorporates 's' for stochastic, and a consonant 'x' to indicate the -arity of the phylogenetic tensor.

Table 2. Comparison of Molien series for binary multitaxon versus multiqubit systems. [a][18] (le Paige 1881, Schwartz 1922); [b][19]; [c][20] (central term $229752q^{52}$); [d][21, 12] (evaluation of the stochastic binary quartet case is in progress).

L	2^L	$h(q)$, L qubits	$h(q)$, L taxa
2	4	$\frac{1}{(1-q^2)}$	$\frac{1}{(1-q)(1-q^2)}$
$3^{a,d}$	8	$\frac{1}{(1-q^4)}$	$\frac{1+q^3}{(1-q)(1-q^2)^3(1-q^4)}$
$4^{b,d}$	16	$\frac{1}{(1-q^2)(1-q^4)^2(1-q^6)}$	———
5^c	32	$\frac{1+16q^8+\cdots\cdots+16q^{96}+q^{104}}{(1-q^4)^5(1-q^6)(1-q^8)^5(1-q^{10})(1-q^{12})^5}$	———

4. Conclusions and outlook

Understanding the phenomenal complexity of molecular data in the setting of phylogenetics provides a rich and challenging arena not only for statistics and bioinformatics, but also for techniques and insights from statistical physics, biophysics and mathematical physics. Applications of group methods in molecular phylogenetics can help model choice and inference, and can even teach us about entanglement! The discrete, digital nature of the biological molecular information has a natural affinity with algebraic and combinatorial methods, and can even be handled via quantum simulation [22].

> "We have been trained to think of physics as the foundation of biology, but it is possible to realize that indeed biology can also be regarded as a foundation for thought, language, mathematics and even physics."
>
> Louis Kauffman, *Biologic*, 2002

Acknowledgments

This paper is based on a contributed talk at Group 32, Prague, July 2018 and draws from joint research with colleagues in the phylogenetics group, University of Tasmania (Jarvis and Sumner 2018 [12]). The work is supported under ARC Discovery project grants.

References

[1] Darwin C R accessed 1/5/18 *Notebook B: [Transmutation of species (1837-1838)]. CUL-DAR121, p.36* (Transcribed by Kees Rookmaaker, Darwin Online, http://darwin-online.org.uk/)
[2] Swofford D L 2003 *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0b10* Sinauer Associates, Sunderland, Massachusetts.
[3] Kimura M 1983 *The neutral theory of molecular evolution* (Cambridge University Press)
[4] Wikid25[CC0] https://commons.wikimedia.org/wiki/file:an_excerpt_of_a_multiple_sequence_alignment_of_tmem66_proteins
[5] Yang Y, Wang Q, Chen Q, Yin X, Qian M, Sun X and Yang Y 2017 *Nature Scientific Reports* **7:5306**
[6] Felsenstein J 2004 *Inferring Phylogenies* (Sinauer Associates, Sunderland)
[7] Steel M 2016 *Phylogeny: Discrete and random processes in evolution* (Philadelphia: SIAM)
[8] J G Sumner, Fernández-Sánchez J and Jarvis P D 2012 *Journal of Theoretical Biology* **298** 16–31
[9] Fernández-Sánchez J, Sumner J G, Jarvis P D and Woodhams M D 2015 *Journal of Mathematical Biology* **70** 855–891

[10] Woodhams M D, Fernández-Sánchez J and Sumner J G 2015 *Systematic Biology* **64** 638–650 (*Preprint* http://sysbio.oxfordjournals.org/content/64/4/638.full.pdf+html) URL http://sysbio.oxfordjournals.org/content/64/4/638.abstract
[11] Jarvis P D and Sumner J G 2016 *Journal of Mathematical Biology* **73** 259–282 ISSN 1432-1416 URL http://dx.doi.org/10.1007/s00285-015-0951-7
[12] Jarvis P D and Sumner J G 2018 Systematics and symmetry in molecular phylogenetic modelling: perspectives from physics (arxiv:1809.03078v2 [q-bio]) Topical review
[13] Jarvis P D and Sumner J G 2014 *The ANZIAM Journal* **56** 105–115
[14] Sumner J G, Charleston M A, Jermiin L S and Jarvis P D 2008 *Journal of Theoretical Biology* **253** 601–615
[15] Holland B R, Jarvis P D and Sumner J G 2013 *Systematic Biology* **62** 78–92
[16] Lake J A 1994 *Proceedings of the National Academy of Sciences* **91** 1455–1459
[17] Sumner J G and Jarvis P D 2006 *Mathematical Biosciences* **204** 49–67
[18] Dür W, Vidal G and Cirac J I 2000 *Physical Review A* **62** 062314
[19] Luque J G and Thibon J Y 2003 *Physical Review A* **67** 042303
[20] Luque J G and Thibon J Y 2005 *Journal of Physics A: Mathematical and General* **39** 371
[21] Hewson T J, Sumner J G and Jarvis P D **In preparation**
[22] Ellinas D and Jarvis P D 2018 arXiv:1105.1582 [quant-ph]